INTERSPEECH 2009 BRIGHTON

A Quantitative Study of F0 Peak Alignment and Sentence Modality

Hansjörg Mixdorff [1] *Hartmut R. Pfitzinger*[2]

[1] Department of Informatics and Media, BHT University of Applied Sciences, Berlin, Germany
[2] Institute of Phonetics and Digital Speech Processing, University of Kiel, Germany
mixdorff@beuth-hochschule.de, hpt@ipds.uni-kiel.de

Abstract

The current study examines the relationship between prosodic accent labels assigned in the *Kiel Corpus of Spontaneous Speech IV,* Isačenko's intoneme classes of the underlying accents and the associated parameters of the Fujisaki model. Among other findings, there is a close connection between early peaks and information intonemes, as well as late peaks and non-terminal intonemes. The majority of tokens within both intoneme classes, however, are associated with medial peaks. Precise analysis of alignment shows that accent command offset times for information intonemes are significantly earlier than for non-terminal intonemes. This suggests that the anchoring of the relevant tonal transition could be more important for separating different intonational categories than that of the F0 peak.
Index Terms: Fujisaki model, prosodic labeling, sentence mode

10.21437/Interspeech.2009-297

1. Introduction

The current study examines the relationship between prosodic accent labels assigned in the *Kiel Corpus of Spontaneous Speech IV,* parameters of the Fujisaki model, as well as the underlying intoneme class and focal condition. Recently, the question was investigated whether alignment differences across languages arise from a continuum of phonetic alignment realizations which fall within a single phonological category [1] or are realizations of several phonological categories [2].

1.1 The Concept of Intonemes and their Quantitative Analysis

In the works of Isačenko and Schädlich [3] and Stock and Zacharias [4], a given *F0* contour is mainly described as a sequence of communicatively motivated tone switches, major transitions of the *F0* contour aligned with accented syllables. Tone switches can be thought of as the phonetic realization of phonologically distinct intonational elements, the so-called "intonemes". In the original formulation by Stock, depending on their communicative function, three classes of intonemes are distinguished, namely the N↑ intoneme ("non-terminal intoneme", signalling incompleteness and continuation, rising tone switch), I↓ intoneme ("information intoneme" at declarative-final accents, falling tone switch, conveying information), and the C↑ intoneme ("contact intoneme" associated, for instance, with question-final accents, rising tone switch, establishing contact). Hence intonemes in the original sense mainly distinguish sentence modality, although there exists a variant of the I↓ intoneme, I(E)↓ which denotes emphatic accentuation and occurs in contrastive, narrowly focused environments. Intonemes for reading style speech are predictable by applying a set of phonological rules to a string of text as to word accentability and accent group formation.

Based on this concept, Mixdorff and Jokisch [5] developed a model of German prosody anchoring prosodic features such as F0, duration, and intensity to the syllable as a basic unit of speech rhythm. In order to quantify the interval and timing of the tone switches with respect to the syllabic grid, the framework adopts the well-known quantitative Fujisaki model for parameterizing *F0* contours [6]. The Fujisaki model reproduces a given *F0* contour by superimposing three components: A speaker-individual base frequency *Fb*, a phrase component and an accent component. The phrase component results from impulse responses to impulse-wise phrase commands associated with prosodic breaks. Phrase commands are described by their onset time *T0*, magnitude *Ap* and time constant *alpha*. The accent component results from step-wise accent commands associated with accented syllables. Accent commands are described by on- and offset times *T1* and *T2*, amplitude *Aa* and time constant *beta*.
In a perception study [7] employing synthetic stimuli of identical wording but varying *F0* contours created with the Fujisaki model it was shown that information intonemes are characterized by an accent command ending before or early in the accented syllable, creating a falling contour. N↑ intonemes were connected with rising tone switches to the mid-range of the subject connected with an accent command beginning early in the accented syllable and plateau-like continuation up to the phrase boundary, whereas C↑ intonemes required *F0* transitions to span a total interval of more than 10 semitones and generally starting later in the accented syllable, although the *F0* interval was a more important factor than the precise alignment.
Mixdorff and Fujisaki [8] compared German ToBI labels with Fujisaki model parameters on a corpus of news reading. They found that tone labels were strongly correlated with accent commands, and the type of label (typically H*L and L*H) was clearly reflected by the onset and offset times of these accent commands. These main label types once again correspond to the I↓- and N↑ intonemes in Stock's formulation, respectively.

1.2 The Kiel Intonation Model and PROLAB

An alternative approach to the symbolic description of German intonation is the *Kiel Intonation Model* (KIM) [9]. The prosodic annotation scheme PROLAB [10] which is based on KIM comprises the functional and phonological distinction between early, medial, and late peaks [11][12].
The concept of the prosodic labelling system PROLAB provides three different pitch peak synchronisations: *early* (F0 maximum located before the accented vowel), *medial* (maximum within the accented vowel), and *late* (maximum late within, or after, the accented vowel). Figure 1 shows three typical examples of an early, medial, and late peak taken from the *Kiel Corpus of Spontaneous Speech IV*. The prosodic labels are ')', '^', and '(', respectively. PROLAB also provides four different accent levels: *reinforced accent*, *default accent*, *partial deaccentuation*, and *complete*

Copyright © 2009 ISCA

deaccentuation symbolized with '3', '2', '1', and '0', respectively. Labellers based their decisions on three criteria: (1) perceptual-phonetic assessment of the local *F0* contour, (2) visual inspection of *F0* values displayed in a window synchronous to the speech signal, and (3) functional-semantic categorization only within the range of a word whether an information is given (early), new (medial), or unexpected (late).

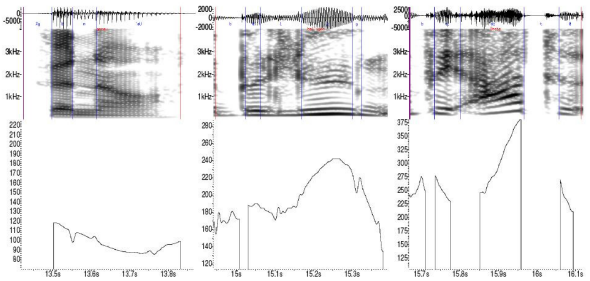

Figure 1: *Examples of an early (left: l05uho51.wav), medial (middle: l02aha16.wav), and late peak (right: l02aha19.wav) taken from the Video Task scenario. From top to bottom: Oscillogram, sonagram, and F0 contour. The y-scales of the F0-panels are adjusted to the F0 range in the respective sound file (left: male speaker, middle/right: female speaker).*

1.3 Quantitative Analysis of PROLAB labels

In a recent study Pfitzinger and Mixdorff parameterized a subcorpus of the *Kiel Corpus of Spontaneous Speech IV* using the Fujisaki model and related accent commands yielded with the underlying PROLAB labels [13]. By applying the Fujisaki model to *F0* contours extracted from spontaneous speech of four speakers *F0* peaks were made accessible to empirical analysis of the temporal alignment with segmental landmarks, e.g. the vowel onset of the accented syllable, and of the relationship between F0 interval and accent level class. A good agreement between PROLAB labelled peak timing and accent command positions relative to the syllable nuclei was found. Early peaks were characterized by an accent command offset time of 41 ms earlier than the accented vowel onset. Medial peaks showed an exact alignment of the accent command center point and the accented vowel onset and possibly an alignment of the accent command offset time with the vowel offset. Late peaks were characterized by an accent command onset time of 18 ms earlier than the accented vowel onset. Accent command amplitudes were significantly different for the three accent levels: A reinforced accent was realized with an *F0* interval of 9.7 semitones, default accents with 5.5 semitones, and partial deaccentuation with 3.9 semitones.

1.4 Focus of the Current Study

The current study is intended to expand the work in [13] towards the following research questions:
If, as it has been shown, the course of the *F0* contour associated with the accented syllable (falling in the case of I↓, rising for C↑ and N↑intonemes) is crucial with regard to signalling sentence modality, how does this function interact with the semantic function postulated by Kohler as being a property of early, medial and late *F0* peak alignment (given, new, unexpected)? Is there some kind of mapping, overlap or correspondence of categories between the two lines of interpretation?
It should be noted that Kohler exclusively demonstrated the functionally-semantic distinction of peak-alignment on examples of short, declarative single-accent utterances [12] in which the accents all correspond to I↓ intonemes in the Stock sense. Falling declarative tone switches can occur very early before the accented syllable, during the syllable or even after it. In this case, we could hypothesize, that the Kohler distinctions might lead to a further subdivision or refinement of the I↓ intoneme class.
The labellers of the *Kiel Corpus*, however, as will be shown later, assigned the peak alignment labels to every intonationally marked accent in the corpus, regardless of its terminal (declarative-final) or non-terminal character.
By labelling the intoneme classes for all accents in the subcorpus we are enabled to investigate their correspondences with PROLAB peak labels. Furthermore, the actual alignment properties can be examined quantitatively based on the temporal and amplitude characteristics of the underlying accent commands.

2. Speech Material and Method of Analysis

The speech material consists of a subcorpus from the *Video Task scenario* (or *Lindenstrasse Daily Soap scenario*) of the *Kiel Corpus of Spontaneous Speech IV* ([14], similar but non-identical video material was presented to two subjects sitting in separate quiet, sound-treated rooms. After the presentation, the subjects discussed differences and similarities of what they had seen and heard. They were not able to see each other and communicated via headphones and microphones placed in front of them. The corpus contains the transliteration and audio files (80 minutes, approx. 13,000 consecutive words) as well as time-aligned segmental and prosodic label files of six overlapping German dialogues (4 female and 2 male speaker pairs). Of the 12 speakers, 2 female and 2 male speakers each with approx. 160 seconds of speech were selected for the current study.
The existing *F0* contours, Fujisaki model parameters and accent command-aligned PROLAB accent labels [13] were augmented by a perceptual classification of intoneme classes. For every accent in the subcorpus it was auditorily determined whether the accent pertained to a word with non-terminal-rising (N↑intoneme), terminal-falling (I↓intoneme) or high-rising, question-like (C↑intoneme) intonation. To this effect, the labeller listened to a context beginning one or two words before the word examined leading up to the end of that word. Furthermore, it was determined whether the accented syllable was connected with a word in broad or narrow focus. To this end, the labeller listened to the entire prosodic phrase to which the word pertained. In the discourse, narrow focus was usually connected with some kind of contrast expressed by the talker.

3. Results of Analysis

Figure 2 displays means and standard deviations of temporal alignments for (from left to right) I↓, N↑ and C↑intonemes. The top row contains the graphs for narrowly focused items and the bottom for broadly focused items. The average alignment and accent command amplitude *Aa* for each case is indicated by the box-shape of the accent command and its resulting smoothed output drawn with respect to the syllable nucleus marked in grey. As can be seen, narrow focus is connected with higher accent command amplitudes than broad focus. At least for the broad focus condition (bottom row) there is a distinction between I↓ (early), N↑ (medial) and C↑ (late) intonemes with respect to their alignment with the syllable nucleus.

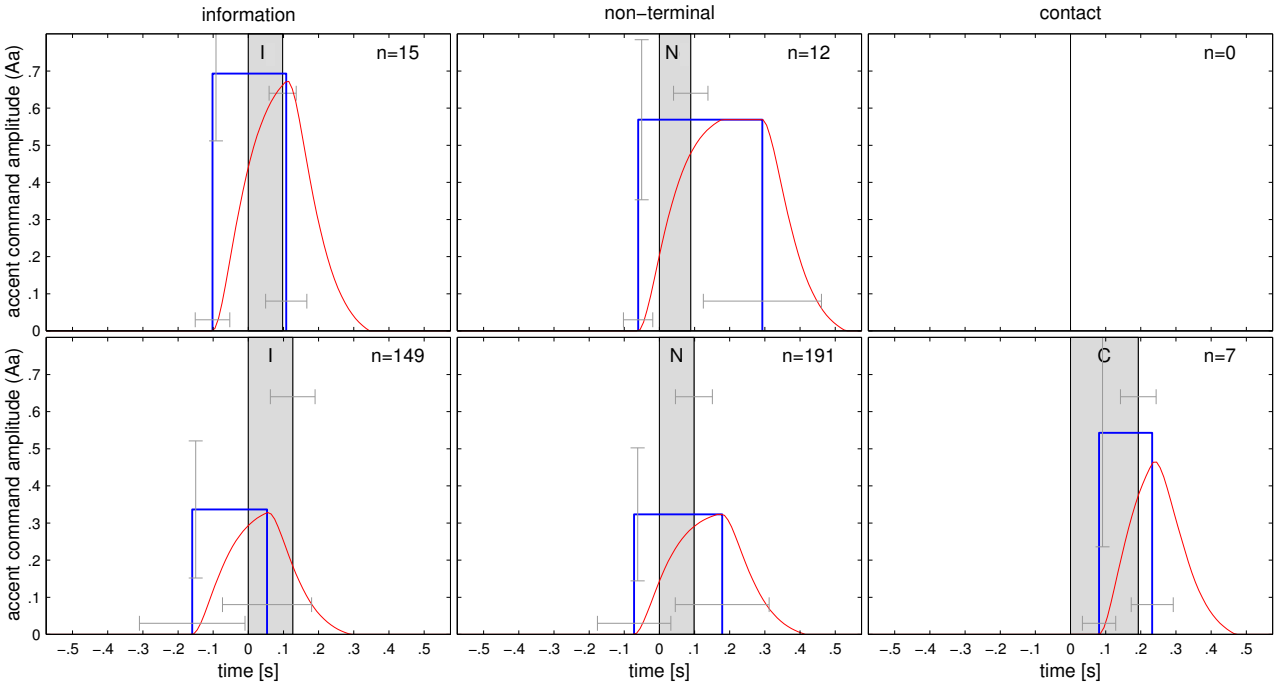

Figure 2: *Means and standard deviations of accent command alignment with respect to the vowel nucleus onset and accent command amplitude Aa for all three intoneme classes. From the left to the right: I↓ intoneme, N↑ intoneme and C↑ intoneme. Top row: Narrowly focused, bottom row: broadly focused. The horizontal whiskers indicate the standard deviations of accent command onset time T1 and offset time T2 and nucleus offset time relative to the onset of the nucleus, and the vertical whisker the standard deviation of accent command amplitude Aa.*

In the narrow focus case there does not seem to be a clear separation between N↑ and C↑intonemes regarding their timing.

Table 1 displays the correspondences between intoneme classes and PROLAB accent labels. The following discussion mostly concentrates on the difference between N↑ and I↓ intonemes for broadly focused items which concern the majority of labels (rows 1 and 2 in the table). As can be seen, most of the I↓ intonemes are associated with either early or medial labels, whereas most of the N↑ intonemes are connected with either medial or late peaks.

Table 1: *Correspondences between PROLAB accent labels and intoneme classes (number of occurrences).*

Intoneme class	PROLAB accent label							
	1)	1^	1(	2)	2^	2(	3^	3(
1: I↓ (broad)	18	11	6	64	69	10	8	1
2:N↑ (broad)	6	34	16	12	84	66	2	7
3:C↑ (broad)	0	0	0	4	1	2	0	0
4:I↓ (narrow)	0	0	0	0	4	1	5	5
5:N↑ (narrow)	0	0	1	0	1	8	2	1

Chi-square test for PROLAB accent levels 1 and 2 confirms that *early peaks* as opposed to *late peaks* are significantly correlated with the I↓ and N↑ intoneme categories, respectively. For accent level 1, the result is very signficant ($\chi 2(1, N=46)=10.48$, $p=.012$, two-sided) and for level 2 highly significant ($\chi 2(1, N=152)=76.79$, $p<.001$, two-sided).

Despite their obvious functional difference, however, the majority of I↓ and N↑intonemes are *associated with medial peaks*.

In the case of narrowly focused items, which is harder to interpret due to the small numbers, we also find a slight preference for earlier alignment in I↓ intonemes (mostly medial peaks) than in N↑intonemes (mostly late peaks). The number of occurrences of C↑intonemes is too small to make any claims about their preferred alignment.

Besides these major findings, the table also shows that narrowly focussed items are generally associated with accent levels 2 and 3 of the PROLAB system.

In order to examine whether the temporal characteristics of N↑ and I↓ intonemes associated with medial peaks are essentially the same, we calculated means and standard deviations of accent command onset time $T1_{rel}$ and accent command offset time $T2_{rel}$. These are measured relative to the vowel nucleus onset time. We included all combinations of I↓ and N↑ intonemes aligned with PROLAB peak assignments, but pooled across accent levels, considering only broadly focused items. Table 2 shows, for instance, that on the average accent commands associated with I↓ intonemes and early peak assignment start 225 ms before the vowel onset and end 9 ms after it. This means that the resulting *F0* contour falls across the vowel nucleus.

The results suggest that in all three conditions (early, medial and late peaks) the accent command in I↓ intonemes occurs earlier than in the N↑ intoneme. The case of medial peaks is of special interest for us because the majority of I↓ and N↑. intonemes falls into this category. Although $T1_{rel}$ is very similar for both cases (-132ms vs. -112ms), $T2_{rel}$ is considerably earlier for the I↓ intoneme than for the N↑ intoneme (65ms vs. 146ms). Independent samples T test indeed shows that this difference is highly significant ($T=4.7$, $df=174$, $p < .001$) whereas it is not for $T1_{rel}$.

Table 2: *Temporal alignment of I↓ and N↑ intonemes.*

peak	intoneme		$T1_{rel}$ [ms]	$T2_{rel}$ [ms]
early	I↓	mean	-225	9
		s.d.	175	145
		N	62	62
	N↑	mean	22	210
		s.d.	100	94
		N	6	6
medial	I↓	mean	-132	65
		s.d.	107	82
		N	72	72
	N↑.	mean	-112	146
		s.d.	105	132
		N	104	104
late	I↓	mean	-53	126
		s.d.	124	124
		N	14	14
	N↑	mean	-20	232
		s.d.	63	125
		N	80	80

If we consider the peak alignment, the N↑ intoneme, in order to signal continuation, requires a rising tone switch (which occurs at the onset of the accent command) and a high value of *F0* at least until the end of the vowel nucleus. This explains the later offset time $T2_{rel}$. In contrast, the I↓ intoneme requires a falling *F0* contour (which starts at the end of the accent command) and can continue right across the following syllable. Another slight indication for stronger anchoring of the *F0* fall in I↓ intonemes is the smaller standard deviation of $T2_{rel}$ compared with that of $T1_{rel}$. The situation is reversed for N↑ intonemes where the rise, i.e. $T1_{rel}$ appears to be more strongly anchored. Similar explanations hold for the case of late peaks. As discussed before, certain I↓ intonemes exhibit late accent command offsets with the *F0* fall occurring in the following syllable.
The fact that the 6 N↑ intonemes with early peak assignment occur even later than those in the medial peak condition suggests that they have been placed in the wrong peak category.

4. Discussion and Conclusions

Our results indicate that peak alignment as labelled in the PROLAB system is at least partly influenced by the underlying intoneme class, i.e. the sentence modality signalled by the accent. As discussed before, declarative-final intonation requires an earlier alignment of an accent command relative to an accented syllable as opposed to the later alignment for non-terminal intonation, in order to realize the falling and rising tone switches, respectively. Hence there is a stronger connection between early peaks and I↓ intonemes, and late peaks and N↑ intonemes. Although we found I↓ and N↑ intonemes in all three peak classes, they clearly differ with respect to their precise alignment as reflected by the temporal properties of the underlying accent commands. These findings, however, raise the question why the assignment "medial" as signalling "new information" was profusely employed by the labellers, despite the obvious functional differences between the non-terminal and the information intonemes. Our results suggest that the anchoring of the relevant tone switches could be more important for separating different intonational categories than that of the *F0* peak.
In future work it remains to be examined which communicative functions the accented words fulfil in the discourse, that is, for instance, in which circumstances early, medial and late I↓ intonemes occur. By analysing the interaction between talkers we might arrive at more clearly separated intonational categories, hence refining the concept of intonemes with respect to their alignment.

4. References

[1] Atterer, M. and Ladd, D.R., "On the phonetics and phonology of segmental anchoring of F0: evidence from German", J. of Phonetics 32, 177-197, 2004.

[2] Niebuhr, O. and Ambrazaitis, G., "Alignment of medial and late peaks in German spontaneous speech", in Proc. of the 3rd Int. Conf. on Speech Prosody. Dresden, 161-164, 2006.

[3] Isačenko, A.V., Schädlich, H.J., "Untersuchungen über die deutsche Satzintonation", Akademie-Verlag, Berlin, 1964.

[4] Stock E., Zacharias, C., "Deutsche Satzintonation", VEB Verlag Enzyklopädie, Leipzig, 1982.

[5] Mixdorff, H. and O. Jokisch, O., "Building an Integrated Prosodic Model of German", Proceedings of Eurospeech 2001, vol. 2, pp. 947-950, Aalborg, Denmark, 2001.

[6] Fujisaki, H. and Hirose, K. "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", J. of the Acoustical Society of Japan (E) 5(4), 233-241, 1984.

[7] Mixdorff, H., Fujisaki, H., "Production and perception of statement, question and non-terminal intonation in German", Proc. of ICPhS, Stockholm,, Vol. 2, pp. 410-413, 1995.

[8] Mixdorff, H. and H. Fujisaki, H., "A quantitative description of German prosody offering symbolic labels as a by-product", Proc. of ICSLP 2000, vol. 2. Beijing, 2000.

[9] Kohler, K. J., "A model of German intonation, "Arbeitsberichte (AIPUK) 25, IPDS, Christian-Albrechts-University, Kiel, 295-360, 1991.

[10] Kohler, K. J., "PROLAB, the Kiel system of prosodic labelling", in Proc. of the XIIIth Int. Congress of Phonetic Sciences, vol. 3. Stockholm, 162-165, 1995.

[11] Kohler, K. J., "Categorical pitch perception", In Proc. of the XIth Int Congress of Phonetic Sciences, vol. 5. Tallinn, 331-333, 1987.

[12] Kohler, K. J., "Terminal intonation patterns in single-accent utterances of German: Phonetics, phonology and semantics", Arbeitsberichte (AIPUK) 25, IPDS, Christian-Albrechts-University, Kiel, 115-185.

[13] Pfitzinger, H.R., Mixdorff, H., Peters, B., "Correspondences between KIM-based symbolic prosodic labels and parameters of the Fujisaki model", Nordic Prosody X. pp. 261-272. Helsinki, 2009.

[14] Kohler, K. J., B. Peters, and M. Scheffers (Eds.), "The Kiel Corpus of Spontaneous Speech IV, German: Video Task Scenario (Kiel-DVD1)", Kiel: IPDS, Christian-Albrechts-University, 2006.